Methods for studying changes lacking a variable

Säily, Tanja

John Benjamins
2018

Säily , T , Nurmi , A & Sairio , A 2018 , Methods for studying changes lacking a variable . in T Nevalainen , M Palander-Collin & T Säily (eds) , Patterns of Change in 18th-century English : A Sociolinguistic Approach . Advances in Historical Sociolinguistics , no. 8 , John Benjamins , Amsterdam , pp. 68-74 . https://doi.org/10.1075/ahs.8

http://hdl.handle.net/10138/300078
https://doi.org/10.1075/ahs.8

5.3 Methods for studying changes lacking a variable

Tanja Säily, Arja Nurmi and Anni Sairio

5.3.1 Introduction

Some linguistic structures and forms do not present an obvious variable to analyse. In this volume, we look at three main types of such cases. Firstly, there are cases where it is reasonably simple to identify the possible variable but unfeasible to retrieve it from an untagged corpus. So, in the case of periphrastic DO, the variable would include all instances with a bare main verb that could be used with DO, called the "simple form" by Ellegård (1953). There are some borderline cases regarding verbs to include, but the main principle is clear, and it is merely the problem of data retrieval that prohibits the use of the variable. (But see Chapter 8 for a more detailed discussion.)

The second type of structure without a variable is one where identifying the form is more problematic. This is illustrated by the progressive, where the varying complexity of the verb phrase associated with the form does not allow for a clear definition: all VPs do not accept the progressive aspect, and a simple normalisation of the frequencies does not indicate how much possibility there is for variation to occur. An added difficulty is the spread of the form into new grammatical contexts over the period studied: should the variable change accordingly? There would also be the choice of treating all finite verb phrases as the variable, but this would again require a tagged corpus. For a discussion of the problems associated with defining the variable for the progressive, see Smitterberg (2005), Kranich (2010), and Chapter 11 below.

The third type of change without a variable concerns a different type of problem, where defining a reasonable variable would not be solved with the help of a tagged corpus. The study of derivational morphemes *-ness* and *-ity* leaves even more problems open for question. Should the variable include all abstract nouns? All nouns with any derivational morpheme? What about the bases: Should we include all adjectives (and other word classes) which could be used as bases for these morphemes? Only those bases where each of the morphemes in question is genuinely possible? For example, *-ity* can generally only be used with bases of a French or Latinate origin, whereas *-ness* takes both foreign and native bases. For further discussion of defining the linguistic variable in derivational morphology, see Cowie (1999: 183–189) and Säily (2014: 33–34).

Owing to the issues outlined above, we abandon the idea of constructing a linguistic variable for three of our changes in this volume: periphrastic DO, the progressive, and the nominal suffixes *-ness* and *-ity* (Chapters 8, 11 and 12 below). Instead of expressing the frequency of these forms as a proportion of the overall frequency of a variable, we can normalise it as a proportion of the number of running words

in the corpus (Nurmi 1996). But how do we know whether the observed change in the frequency is statistically significant? Recent research (e.g. Kilgarriff 2001; Lijffijt et al. 2012; Bestgen 2014) has shown that many widely used tests of statistical significance, such as the chi-square and log-likelihood ratio tests, are in fact inappropriate for comparing word frequencies in texts, because they assume that all words occur independently from each other, which is never true. Moreover, change is often visualised using simple line graphs, which hide the variability within the corpus.

5.3.2 Method 1: accumulation curves and permutation testing

We solve the problems of visualising change and determining its statistical significance by using two robust methods. The first of these was proposed by Säily & Suomela (2009). Involving accumulation curves and the statistical technique of permutation testing, this method assesses significance without resorting to simplifying assumptions. It was originally developed for comparing type frequencies in the study of morphological productivity, as in our third type of change above. *Type frequency* refers to the number of different words formed using the element under study in the corpus, whereas *token frequency* refers to the number of all occurrences of the words under study in the corpus.

Unlike token frequencies, type frequencies present the additional complication that they cannot be normalised, because normalisation presupposes that the growth rate of the frequency is linear, which is not the case with type frequency. As corpus size increases, the number of types increases in a nonlinear manner, with more new types being encountered when the corpus is small, and the growth rate decreasing as the corpus gets larger. Therefore, type frequencies obtained from corpora of different sizes, such as subcorpora representing different time periods, cannot be compared through normalisation. Our method eliminates the need to normalise, and it can be applied to both type and token frequencies.

The idea behind the method is as follows. Instead of trying to compare subcorpora of different sizes, we compare actual subcorpora with randomly composed subcorpora of the same size. The randomly composed subcorpora are obtained by dividing the entire corpus into samples and by randomly sampling these using a statistical technique called permutation testing. The samples need to be large enough to preserve discourse structure; they can consist of individual texts or e.g. all texts written by a person during a time period. For each corpus size, a million random subcorpora are sampled by a computer program (Suomela 2014). These random subcorpora give upper and lower bounds for the type or token frequency of the form in question. If the frequency observed in the actual subcorpus is higher than, say, 99.9% of the observations in random subcorpora of the same size, we can say that the frequency is significantly high at a probability of $p < 0.001$. For a more detailed description of the method, see Säily (2014).

Figure 5.2 illustrates the results of applying the method to type frequencies of the suffix *-ity* in the 17th-century section of the CEEC, 1600–1681. The *x*-axis shows corpus size in running words, while the *y*-axis shows type frequency. The shaded areas display the range of type frequencies of *-ity* in the random subcorpora: the darkest area in the middle covers most of the subcorpora, and when the next darkest area is added, the coverage increases to 90%, then 99%, 99.9% and finally 99.99% (pure white). We can see that type frequency increases with corpus size in a nonlinear manner: the shaded plot is curved, not straight. We call these figures "type accumulation curves", although "cucumiform plot" has also been suggested to describe the shape. If we look at the actual subcorpus of letters written during the first 40-year period in the corpus, 1600–1639, we can see that its type frequency falls in the lightest grey area. This means that fewer than 0.1% of the randomly composed subcorpora of the same size have such a low type frequency, making the productivity of *-ity* significantly low in this period at $p < 0.001$. The second period, 1640–81, is not significantly different from randomly composed subcorpora of the same size, as up to 10% of them have a similarly high type frequency ($p < 0.1$). From these results we may deduce that the productivity of *-ity* increases over time in the corpus.

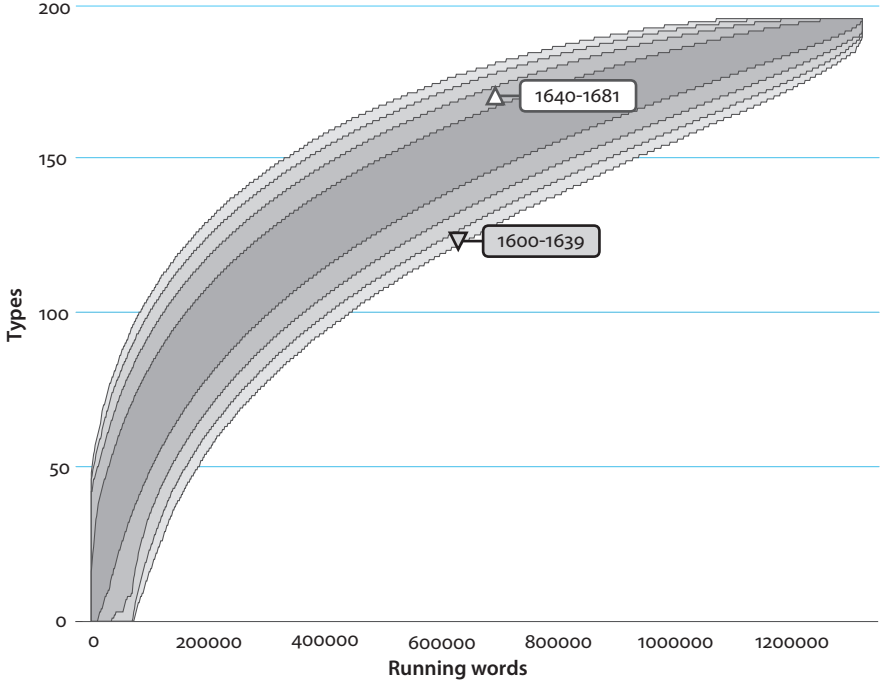

Figure 5.2 Bounds for the type frequency of the suffix *-ity* in the 17th-century section of the CEEC, 1600–1681

Figure 5.3 demonstrates the use of the method with token frequencies, namely those of periphrastic DO in affirmative statements in the 18th-century section of the CEEC, 1680–1800. Note that token frequency does grow linearly with corpus size, resulting in a straight rather than curved plot. Nevertheless, we call these figures "token accumulation curves"; the term "cucumiform plot" is also applicable here as cucumbers can be either curved or straight. We can see that the first 40-year period, 1680–1719, has a significantly high token frequency, as practically all of the randomly composed subcorpora of the same size have a lower token frequency than it ($p < 0.0001$). The middle period of 1720–1759 is not significantly different from random subcorpora, but the last period, 1760–1800, uses DO significantly less frequently than random subcorpora of the same size ($p < 0.001$). We may thus state that the use of affirmative DO declines significantly over time in this corpus (see further Chapter 8 below).

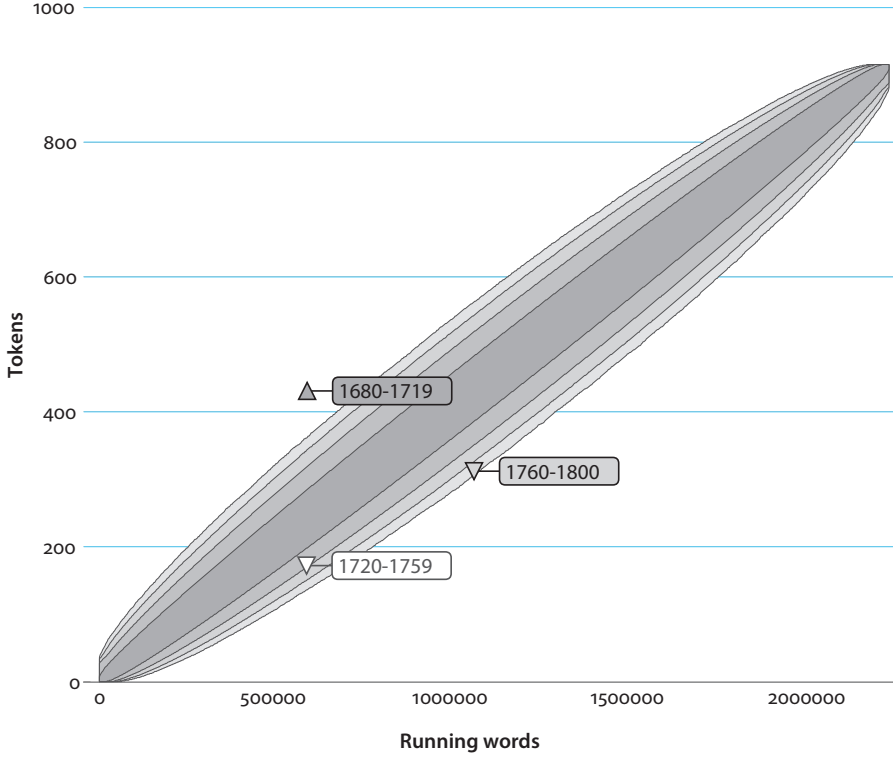

Figure 5.3 Bounds for the token frequency of periphrastic DO in affirmative statements in the 18th-century section of the CEEC, 1680–1800

5.3.3 Method 2: beanplots and the Wilcoxon rank-sum test

While the accumulation curves technique is a marked improvement over many earlier methods, it does have some drawbacks. Firstly, as a measure of significance it can sometimes be overly conservative (Lijffijt 2013: 35, 38), so that some genuine differences may be classified as non-significant. Secondly, the visualisation is perhaps unintuitive for studying change over time, as the *x*-axis displays corpus size rather than time. Therefore, it is here complemented with another method applicable to token frequencies (in our case, affirmative DO and the progressive). First applied to historical sociolinguistics by Vartiainen et al. (2013), this method visualises change using beanplots (Kampstra 2008) and assesses statistical significance using the Wilcoxon rank-sum test, also known as the Mann–Whitney U test (Wilcoxon 1945; Mann & Whitney 1947).

Some scholars in diachronic corpus linguistics have been moving from line graphs to boxplots because the latter provide more information on the variability within subcorpora. As a further improvement, Säily et al. (2011) introduce the beanplot to the field (for a comparison of boxplots and beanplots, see Säily 2014: 57–59). Figure 5.4 presents a beanplot view of change in affirmative DO, which was illustrated using accumulation curves in Figure 5.3. The *x*-axis shows time period, while the *y*-axis shows normalised token frequency. In the middle of each of the three "beans" is a vertical scatterplot, where each thin tickmark represents the normalised frequency of affirmative DO in one sample; the tickmarks give an indication of the variability and amount of data within the time period. The samples consist of a person's letters to a specific kind of recipient (nuclear family, other family, family servants, close friends or other acquaintances) during a 20-year period. This allows us to study the change in sufficient detail while ensuring that no one person contributes more than a few samples and thus cannot easily skew the results. The thicker line going horizontally through each bean represents the median frequency of the samples – while the original beanplot (Kampstra 2008) uses mean frequency, Vartiainen et al. (2013) change this to median, noting that the latter is more robust to outliers. The shape of the bean reveals the distribution of the samples, which is crucial to identifying outliers. Here the shape is mirrored on either side of the scatterplot, but beanplots can also be formed so that they consist of two different subcorpora, e.g. women on the left and men on the right, allowing for easy comparison between the two.

The beanplot shows that the normalised frequency of affirmative DO decreases over time in the corpus, the median practically dropping to zero after the first period. To assess whether the decrease between the first two periods is statistically significant, we use the Wilcoxon rank-sum test (Wilcoxon 1945; Mann & Whitney 1947). Like permutation testing, this test is assumption-free; furthermore, Lijffijt

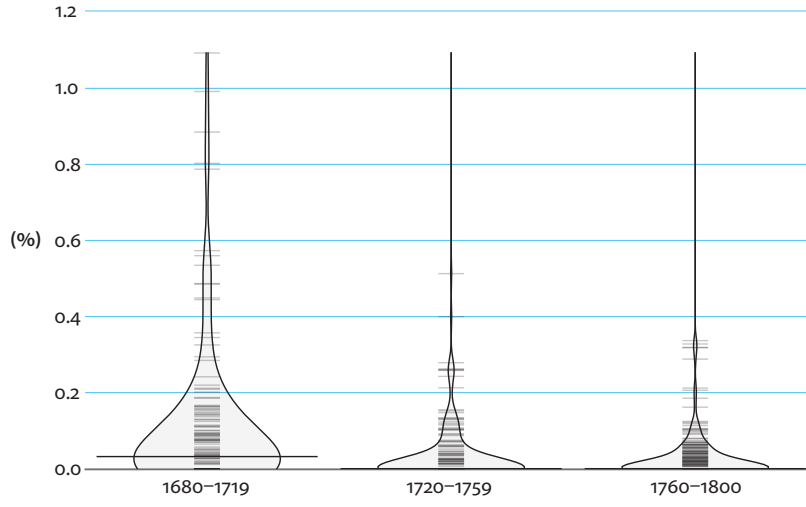

Figure 5.4 Beanplot: normalised frequencies of periphrastic DO in affirmative statements in the 18th-century section of the CEEC, 1680–1800

et al. (2016) find it to be among the best-performing methods for comparing word frequencies. We formulate the hypothesis that the normalised frequency of affirmative DO is smaller in samples produced in the period 1720–1759 than in 1680–1719. We then make a list of the samples from both of these periods and order the list by normalised frequency. The null hypothesis is that the samples from both periods are distributed evenly within the ordered list. The test measures how surprising the actual distribution is compared to the null hypothesis, and the p-value produced by the test tells us the probability that we are wrong in rejecting the null hypothesis. In this case, $p \approx 0.000000$, so the probability is extremely low. We may thus say that the use of affirmative DO declines significantly between these two periods in the corpus.

5.3.4 Addendum: multiple hypothesis testing

There is an additional component to both of the methods presented above. The p-value yielded by significance testing indicates the probability that we are wrong in rejecting a single null hypothesis. However, we need to test multiple hypotheses, as we wish to compare many social categories and time periods. The more hypotheses we test, the greater the probability that we are wrong in rejecting the null hypothesis in at least some of the cases. Therefore, we need to adjust the significance level (which is conventionally set to $p < 0.05$) to reflect the number of hypotheses tested. There are many ways to do this; we have chosen to control the **false discovery rate**, or the proportion of false positives out of all positives, using

Benjamini & Hochberg's (1995) procedure. For a simple description of the procedure, see Säily (2014: 50–51). We have chosen an acceptable false discovery rate of 10%, which leads to a different significance level for each of our changes and methods, depending on the number of hypotheses tested and the *p*-values gained. For instance, in the case of affirmative DO and the Wilcoxon rank-sum test (194 hypotheses tested), the significance level is $p < 0.002$. In the chapters analysing our changes, we will use these adjusted significance levels.

Acknowledgments

The authors would like to thank Jukka Suomela for methodological assistance and CSC – IT Centre for Science, Finland, for computational resources.